One-way ANOVA for feature selection

Problem

For the dataset shown below, find the significance of each feature by computing the F-statistic values.

S. No.	Temperature	Humidity	Wind	Match Played
1	30	70	10	No
2	25	60	15	No
3	26	65	12	No
4	28	75	9	No
5	29	50	11	No
6	22	85	5	Yes
7	18	90	7	Yes
8	24	80	14	Yes
9	21	95	6	Yes
10	23	55	8	Yes

Solution

We have two groups representing match conditions:

- **Group 1**: "Match played: No"
- **Group 2**: "Match played: Yes"

The data includes three variables:

- **Temperature (Column 1)**
- **Humidity (Column 2)**
- **Wind (Column 3)**

Our goal is to conduct a one-way Analysis of Variance (ANOVA) for each variable to determine if there is a significant difference in means between the two groups. For the classification to be 'great', the variation of values within each group should be less and the variation of the means of values between

groups. This implies that the ratio of variation of means of values between groups and sum of the variation of values within each group, should be greater for the classification to be 'great'. F-statistic is a metric that signifies precisely this ratio.

Data for Each Variable

1. **Temperature (Column 1)**: - Group 1: $[30, 25, 26, 28, 29]$ - Group 2: $[22, 18, 24, 21, 23]$
2. **Humidity (Column 2)**: - Group 1: $[70, 60, 65, 75, 50]$ - Group 2: $[85, 90, 80, 95, 55]$
3. **Wind (Column 3)**: - Group 1: $[10, 15, 12, 9, 11]$ - Group 2: $[5, 7, 14, 6, 8]$

Common Values

- $n_{\text{No}} = 5$: number of observations in Group 1.
- $n_{\text{Yes}} = 5$: number of observations in Group 2.
- $N = n_{\text{No}} + n_{\text{Yes}} = 10$: total number of observations.

Column 1: Temperature

For **Temperature**, we have: - Group 1: $[30, 25, 26, 28, 29]$ - Group 2: $[22, 18, 24, 21, 23]$

Step 1: Calculate Group Means and Overall Mean

1. **Group Means**:

$$\bar{X}_{\text{T, No}} = \frac{30 + 25 + 26 + 28 + 29}{5} = 27.6$$

$$\bar{X}_{\text{T, Yes}} = \frac{22 + 18 + 24 + 21 + 23}{5} = 21.6$$

2. **Overall Mean**:

$$\bar{X}_{\text{T}} = \frac{(30 + 25 + 26 + 28 + 29) + (22 + 18 + 24 + 21 + 23)}{10} = 24.6$$

Step 2: Calculate SSB

Sums of Squares Between Groups (SSB):

$$\begin{aligned}\text{SSB}_{\text{T}} &= n_{\text{No}}(\bar{X}_{\text{T, No}} - \bar{X})^2 + n_{\text{Yes}}(\bar{X}_{\text{Yes}} - \bar{X})^2 \\ &= 5 \times (27.6 - 24.6)^2 + 5 \times (21.6 - 24.6)^2 \\ &= 5 \times 9 + 5 \times 9 = 90\end{aligned}$$

Step 3: Calculate SSW

Sums of Squares Within Groups (SSW):

1. **For Group 1**:

$$(30 - 27.6)^2 + (25 - 27.6)^2 + (26 - 27.6)^2 + (28 - 27.6)^2 + (29 - 27.6)^2 = 17.2$$

2. **For Group 2**:

$$(22 - 21.6)^2 + (18 - 21.6)^2 + (24 - 21.6)^2 + (21 - 21.6)^2 + (23 - 21.6)^2 = 21.2$$

Total **SSW** for Temperature:

$$\text{SSW}_\text{T} = 17.2 + 21.2 = 38.4$$

Step 4: Calculate Mean Squares (MSB and MSW)

Mean of Squares Between Groups (MSB) = $\frac{SSB}{\text{dof}_\text{SSB}}$

Mean of Squares Within Groups (MSW) = $\frac{SSW}{\text{dof}_\text{SSW}}$

Computation of degrees of freedom (dof)

Let us say there are n total values and k total groups. While computing SSB, we would compute the variation of mean of each class from the overall mean. This seems as if the mean of each class is a free parameter, thereby making it appear as if there are k free parameters. But the overall mean is computed before computing these variations, i.e., the overall mean is fixed. This means that if the means of $k - 1$ classes are known then the mean of the remaining class would already be determined, because the overall mean is fixed. Hence, the effective number of free parameters is not k but rather $k - 1$. Hence, we have

$$\text{dof}_\text{SSB} = k - 1$$

Likewise, while computing total SSW, we would compute the variation of mean of each parameter with respect to each value of each class. It appears as if the number of free parameters here is n. However, while computing the variation of each value in a group, the mean of that group needs to be computed prior to computation of variation. In other words, the mean of each group is predefined, thereby reducing the number of free parameters. Since we have k groups, there would be k mean values predefined. Hence, the number of free parameters would only be $n - k$.

$$\text{dof}_\text{SSW} = n - k$$

For the current problem, $n = 10$ and $k = 2$. Hence, $\text{dof}_\text{SSB} = 1$ and $\text{dof}_\text{SSW} = 8$[1].

[1] These dof values are common for all features.

Therefore, we have

$$\text{MSB}_\text{T} = \frac{90}{1} = 90, \quad \text{MSW}_\text{T} = \frac{38.4}{8} = 4.8$$

Step 5: Calculate F-statistic

The F-statistic value for Temperatre column is given by

$$F_\text{T} = \frac{\text{MSB}_\text{T}}{\text{MSW}_\text{T}} = \frac{90}{4.8} = 18.75$$

Column 2: Humidity

For **Humidity**, we have: - Group 1: $[70, 60, 65, 75, 50]$ - Group 2: $[85, 90, 80, 95, 55]$

Step 1: Group Means and Overall Mean

1. **Group Means**:

$$\bar{X}_{\text{H, No}} = \frac{70 + 60 + 65 + 75 + 50}{5} = 64$$

$$\bar{X}_{\text{H, Yes}} = \frac{85 + 90 + 80 + 95 + 55}{5} = 81$$

2. **Overall Mean**:

$$\bar{X}_H = \frac{(70 + 60 + 65 + 75 + 50) + (85 + 90 + 80 + 95 + 55)}{10} = 72.5$$

Step 2: SSB

$$\begin{aligned}\text{SSB}_\text{H} &= 5 \times (64 - 72.5)^2 + 5 \times (81 - 72.5)^2 \\ &= 5 \times 72.25 + 5 \times 72.25 = 722.5\end{aligned}$$

Step 3: SSW

1. **Group 1**:

$$(70 - 64)^2 + (60 - 64)^2 + (65 - 64)^2 + (75 - 64)^2 + (50 - 64)^2 = 370$$

2. **Group 2**:

$$(85 - 81)^2 + (90 - 81)^2 + (80 - 81)^2 + (95 - 81)^2 + (55 - 81)^2 = 970$$

Total **SSW** for Humidity:

$$\text{SSW}_\text{H} = 370 + 970 = 1340$$

Step 4: Mean Squares

$$\text{MSB}_\text{H} = \frac{722.5}{1} = 722.5, \quad \text{MSW}_\text{H} = \frac{1340}{8} = 167.5$$

Step 5: F-statistic

$$F_\text{H} = \frac{\text{MSB}_\text{H}}{\text{MSW}_\text{H}} = \frac{722.5}{167.5} = 4.31$$

Column 3: Wind

For **Wind**, we have: - Group 1: $[10, 15, 12, 9, 11]$ - Group 2: $[5, 7, 14, 6, 8]$

Step 1: Group Means and Overall Mean

1. **Group Means**:

$$\bar{X}_{\text{W, No}} = \frac{10 + 15 + 12 + 9 + 11}{5} = 11.4$$

$$\bar{X}_{\text{W, Yes}} = \frac{5 + 7 + 14 + 6 + 8}{5} = 8$$

2. **Overall Mean**:

$$\bar{X}_\text{W} = \frac{(10 + 15 + 12 + 9 + 11) + (5 + 7 + 14 + 6 + 8)}{10} = 9.7$$

Step 2: SSB

$$\begin{aligned}\text{SSB}_\text{W} &= 5 \times (11.4 - 9.7)^2 + 5 \times (8 - 9.7)^2 \\ &= 5 \times 2.89 + 5 \times 2.89 = 28.9\end{aligned}$$

Step 3: SSW

1. **Group 1**:

$$(10 - 11.4)^2 + (15 - 11.4)^2 + (12 - 11.4)^2 + (9 - 11.4)^2 + (11 - 11.4)^2 = 21.2$$

2. **Group 2**:

$$(5 - 8)^2 + (7 - 8)^2 + (14 - 8)^2 + (6 - 8)^2 + (8 - 8)^2 = 50$$

Total **SSW** for Wind:

$$\text{SSW}_\text{W} = 21.2 + 50 = 71.2$$

Step 4: Mean Squares

$$\text{MSB}_\text{W} = \frac{28.9}{1} = 28.9, \quad \text{MSW}_\text{W} = \frac{71.2}{8} = 8.9$$

Step 5: F-statistic

$$F_{\mathrm{W}} = \frac{\mathrm{MSB}_{\mathrm{W}}}{\mathrm{MSW}_{\mathrm{W}}} = \frac{28.9}{8.9} = 3.25$$

Result and Conclusion

$F_{\mathrm{Temperature}} = 18.75$
$F_{\mathrm{Humidity}} = 4.31$
$F_{\mathrm{Wind}} = 3.25$

From the above results, it can be concluded that the feature 'Temperature' has the highest significance, the feature 'Humidity' has a lesser significance, and the feature 'Wind' has the least significance.

Therefore, if only two features are to be selected, then it would be safe to drop the Wind column and work with the first two columns (i.e., Temperature and Humidity), as those two columns have higher F-statistic value.